Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v7n2p254

Comparing two mean humidity curves using functional t-tests: Turkey case

By Keser I.K.

Published: 14 October 2014

Electronic Journal of Applied Statistical Analysis
Vol. 07, Issue 02, October 2014, 254-278
DOI: 10.1285/i20705948v7n2p254

Comparing two mean humidity curves using functional t-tests: Turkey case

İstem Köymen Keser*

Dokuz Eylül University Faculty of Economics and Administrative Sciences Department of Econometry Dokuzçeşmeler Campus 35160 Buca-İzmir / Turkey

Published: 14 October 2014

Functional t-tests involve testing for differences in functional means across two curve groups. If a significant overall difference in the mean curves is detected, one way to identify the location of these differences is pointwise testing. Ramsay and Silverman (2005) suggest using a pointwise test approach based on a permutation method and Cox and Lee (2008) based on Westfall-Young approach. Since both of them have their caveats, here it is suggested to use them simultaneously to improve the validation of results. It is concluded that they strongly support each other. The flexibility of functional data analysis is shown by using basis functions comparatively and the irregular behavior of the differences or mean functions across the years are presented via rainbow plots explicitly. Meteorological data like humidity is used for these aims. Functional t-tests are used in order to observe if humidity mean functions of coastal area and hinterland of Turkey are statistically different. On the other hand, analysis for eleven years (2000-2010) is made in order to observe if these expectations changes in years. Suggested usage of the tests is proven to be practically useful.

keywords: Functional data analysis, functional t-tests, Fourier series, meteorological data, rainbow plots.

1 Introduction

Thanks to the development in technology, the increase in data storage and processing capacities of computers allows storage and analysis of big data, collecting many points

*Corresponding author: istem.koymen@deu.edu.tr

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

on each subject. As an observed intensity in a specific point occurs, it is assumed that the data are in fact sampled from an underlying smooth function. Discussing data as smooth functions caused development of new statistical methods as alternatives to classical statistical methods under the name of Functional Data Analysis. The term "Functional Data Analysis" is firstly defined by Ramsay and Dalzell (1991). The basic philosophy of functional data analysis is approaching observed data functions not as a line of consecutive individual observations, but as unique inputs. In other words, a curve or function is used as the basis unit in data analysis. At the same time, data are often assumed to be a function of time, but this is not a must. As data are made of functions, being able to examine them visually is informative about the structure of data besides appropriate statistical analyses. For this reason, graphical displays are used often in the analysis and interpretation of functional data.

There are many studies in the literature about functional data analysis. Most of these studies are based on the functional expansion of statistical and multivariate statistical methods such as principal components analysis (Besse and Ramsay, 1986, Barra, 2004, Benko, 2004, Lober and Villa, 2004, Ingrassia and Costanzo, 2005, Hall and Nasab, 2006, Keser, 2010), canonical correlation analysis (Leurgans et al., 1993, He et al., 2000, He et al., 2003, He et al., 2004, Kupresanin, 2008), cluster analysis (Cerioli et al., 2005), regression analysis (Ratcliffe et al., 2002, Chiou and Müller, 2007, Ainsworth et al., 2011) and generalized linear models (James, 2002). Besides that, studies on hypothesis tests such as functional t-tests and functional analysis of variance (Cuevas et al., 2004, Shen and Faraway, 2004, Hall and Keilegom, 2007, Cox and Lee, 2008, Kaziska, 2011, Vsevolozhskaya et al., 2013), which are used for testing the equality of two or more mean functions, have started to increase in recent years. All of these methods have been applied on a wide range of areas including the analysis of medical and environmental data, meteorological, financial, econometric and psychological data. On the other hand, Ullah and Finch (2013) provide a detailed literature review about the functional data analysis' field of application especially after 2005.

The goal of this study is to show the usefulness of functional data analysis on real data when classical statistical methods are insufficient. Meteorological data like humidity are used for this aim. It is expected that the effects of geographical location and landform of an area on humidity curve are revealed through functional data analysis. For this aim, Turkey, which has a variable geographical structure and landform, is investigated. The rough terrain, location and position of mountains, seas around the land and increase of height from west to east cause huge climate changes in short distances. Because of these effects, temperature, humidity, and rainfall vary significantly according to regions. Especially positions of regions according to sea, positions of mountains and elevations have significant effect on humidity. This is why, it is expected that humidity functions of coastal area and hinterland will behave different.

Many classical statistical methods are insufficient to test these expectations, because the dimension of the time points (variables) will often be much larger than the sam-

ple size. Thus most of the inferential methods from classic statistical and multivariate analysis cannot be used directly, since they require inversion of the sample covariance matrix. The functional data analysis methods overcome this issue.

In this study, in order to observe if humidity mean functions of coastal area and hinterland are statistically different and where the differences occur, the pointwise tests proposed by Ramsay and Silverman (2005) and Cox and Lee (2008) are used simultaneously. The pointwise approaches used by Ramsay and Silverman (2005) and Cox and Lee (2008) can be considered as follow-up tests to an overall test and both techniques have their caveats. Ramsay and Silverman fail to account for multiplicity issue while performing tests across the evaluation (here time) points. Cox and Lee account for multiplicity but their method cannot assess overall significance (Vsevolozhskaya et al., 2013). Thus in this study it is proposed to use the two methods together to improve the validity of results.

On the other hand, analysis for ten years is made in order to observe if these expectations changes in years. Monthly analyses (e.g. August) are also conducted for the regions with big difference. Besides, Fourier and B-Spline basis approaches are compared. Rainbow plots, which are one of the graphical presentations, are used to interpret changes in years. Additionally, functional t-test results for 11 years are presented graphically and interpreted.

This paper is organized as follows. In 2, basis function approach for the transformation of data observed in discrete points into functional data is introduced. t-tests for comparing means by using permutation tests and Westfall-Young approach for multiplicity are conducted. In 3, individual and mean humidity curves of coastal areas and hinterlands for 35 cities in Turkey for 2010 will be compared in order to observe the effects of geographic location and landform on humidity. The utilization of Fourier and B-Spline basis functions are compared meanwhile. Pointwise methods are used in order to analyze if there are statistical differences among them, where this differences are most dense, and results of the analyses are interpreted statistically and visually. On the other hand, in order to observe if there is a change in humidity structures between 2000 and 2010, mean functions of coastal areas and hinterlands are observed with rainbow plots and functional t-test results are visually presented and summarized. 4 concludes the study and gives further insights.

2 Statistical Methods: Functional Data Analysis

2.1 Basis Function Approach

Flexible methods are needed while estimating functions from the original discrete observations y_i.

$$y_i = x(t_i) + \varepsilon_i \tag{1}$$

A system made of K number of basis functions are chosen for this. The desired $x(t)$ function can be expressed as

$$x(t) = \sum_{i=1}^{K} c_i B_i(t), \;\; i = 1, 2, \ldots, K \tag{2}$$

as a weighted sum of these basis functions. Here, $B_i(t)$ is the i-th basis function and c_i is the corresponding coefficient. c_i coefficients determine the shape and form of the function and can be interpreted as parameters. The method of estimating $x(t)$ as a weighted sum of K basis functions is referred to as Basis Function Approach.

Various basis functions can be used such as B-Splines, polynomials, wavelets and Fourier basis according to the structure of data. Many functions need to repeat itself along a period T. Among the basis functions, Fourier basis function is the most appropriate one for long time periodical data because of its sinusoidal structure, e.g. 11 years. Fourier series are used to model periodical functions and can be useful for periodical data which has a known and fixed period. For $w = \frac{2\pi}{T}$ and the period (T), for a fixed frequency of w, basis can be written as 1, $sin(wt)$, $cos(wt)$, $sin(2wt)$, $cos(2wt)$, ... In other words, after the first fixed basis function, Fourier basis functions are arranged as successive sine/cosine pairs. Functions that are obtained through linear combination of Fourier basis are infinitely differentiable.

B-Spline j with a degree d and order m can be given with a recursive relation as in (3)

$$B_{j,d}(t) = \frac{t - t_j}{t_{j+d} - t_j} B_{j,d-1}(t) + \frac{t_{j+1+d} - t}{t_{j+1+d} - t_{j+1}} B_{j+1,d-1}(t) \tag{3}$$

Generally the knots t are taken as evaluation points. A B-Spline in the form of $B_{j,d}(t)$ depends only on knots $(t_k)_{k=j}^{j+d+1}$ and the number of basis functions is calculated as

$$\text{number of basis functions} = \text{number of knots} + \text{order} - 2 \tag{4}$$

In this study Fourier basis and B-Spline basis are used comparatively to indicate the flexibility of functional data analysis. Since in Fourier basis approach the number of basis functions can be lower than the number of evaluation points, the data can be smoothed with least squares method; so, individual humidity curves that are smooth and infinitely differentiable can be obtained. In B-Spline approach penalized least squares is used to estimate the c_i coefficients.

2.2 Functional t-Test

Functional data analysis gives opportunity to compare curve groups statistically besides revealing the variability and relations in data structures. Functional t-tests and functional analysis of variance are the functional expansions of classical statistical tests such as t-tests and analysis of variance. Development of statistical tests in functional data is

still a vivid area of research.

Just as with ordinary analysis of variance and t-tests, after deciding that the means from two or more samples are significantly different by an overall testing method, one wants to identify more specifically where the differences are. In the functional analysis of variance and functional t test settings, a natural goal is to determine the specific region of t where the differences occur Cox and Lee (2008).

For this problem we can use pointwise testing procedures. Ramsay and Silverman (2005) and Ramsay et al. (2009) propose a pointwise testing procedure based on a permutation method and Cox and Lee (2008) based on a randomization method due to Westfall and Young (1993). Because there are some lacks of both methods we suggest using them simultaneously to improve the validity of results. Both can be used as a follow up test to an overall test to determine the specific region of t where the differences occur. These methods are explained respectively in 2.2.1 and 2.2.2.

In order to test if mean curves of two groups are statistically different, it is assumed that we have two curve groups which are observed in the same time point but may have different number of individual curves.

In other words, $x_{11}(t), x_{12}(t), \ldots, x_{1n_1}(t)$ and $x_{21}(t), x_{22}(t), \ldots, x_{2n_2}(t)$ are assumed to be random samplings that are chosen respectively from $x_1(t)$ and $x_2(t)$ on the same time points t. In order to test if there is a difference between mean curves of two groups , in other words, to compare the means of the two curve groups over time, hypothesis are assumed to be as follows:

$$H_0 : \mu_1(t) = \mu_2(t) \text{ for all } t \in T, \;\; H_1 : \mu_1(t) \neq \mu_2(t) \text{ for at least one } t \in T \tag{5}$$

In both approach firstly test statistic T is used.

$$T(t) = \frac{|\overline{x}_1(t) - \overline{x}_2(t)|}{\sqrt{\frac{1}{n_1}Var(x_1(t))} + \frac{1}{n_2}Var(x_2(t))} \tag{6}$$

It is the functional form of test statistics and $\overline{x}_1(t)$ and $\overline{x}_2(t)$ are the curve means:

$$\overline{x}_1(t) = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}(t), \;\; \overline{x}_2(t) = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}(t) \tag{7}$$

$Var(x_1(t))$ and $Var(x_2(t))$ are the curve variances

$$Var(x_1(t)) = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1i}(t) - \overline{x}_1(t)), \;\; Var(x_2(t)) = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{2i}(t) - \overline{x}_2(t)) \tag{8}$$

for each curve group. As can be seen obviously here, test statistics $T(t)$ is a function of time. After the data of two populations are sampled, $T(t)$ can be calculated and hypothesis test can be practiced.

2.2.1 Functional t-tests based on a permutation method

It is often impossible to derive null distribution of a test statistics. This is especially true in functional data analysis studies in which data is highly or infinitely dimensional. In such cases, permutation tests (or randomization) are used in order to determine the null distribution of $T(t)$ (Lee (2005), Ramsay et al. (2009), Coffey and Hinde (2011)).

A permutation method can be used for obtaining a null distribution for any test statistic and a probability value (p value) to determine the result of the test. This method is based on randomly changing curve labels and calculating test statistics each time. This statistics is referred to as permuted test statistics. This process is repeated ten thousand times in order to obtain a null distribution and it gives a reference in order to evaluate maximum of $T(t)$. At the same time, test statistics are also calculated from original data. Obtained statistics is referred to as observed test statistics. A p-value is obtained by permuted test statistics' ratio that is higher than or equal to observed test statistics. It is assumed that null hypotheses will be rejected for high values of test statistics (Lee, 2005, Ramsay et al., 2009, Coffey and Hinde, 2011).

This procedure is advantageous as it is distribution independent. On the other hand, it is an exact level α test because of the features of permutation test, so, it gives valid p values (Lee, 2005).

Functional t-test that is carried out through permutation tests used by Ramsay and Silverman (2005) and Ramsay et al. (2009) can be summarized with these basic steps:

1. Number of time points (n) and number of permutations (d) are determined.
2. Labels of curves are mixed d times and $T(t)$ is calculated each time for n time points. The results are called as $T_{nullvalues}(t)$ and it is an $n \times d$ dimensional matrix.
3. Maximum differences are taken into consideration for each permutation along with n time points. These differences create a null distribution which is represented with $T_{null}(t)$. As a result, a column vector is obtained at the length of d.
4. For original data, (before changing labels) $T(t)$ is calculated at n time points. This data creates the observed curve. The biggest one of these values is $T_{maxobservation}(t)$. (Observed curve is shown with solid blue line in Figure 4)
5. The point wise critical value is the curve formed by the quantile $(T_{nullvalues}(t), 0.05)$ over all vector permutations at each time point and it is the max value on all of the permutations, so it is a column vector of n dimensions. (It is shown with dashed blue line in Figure 4.)
6. Maximum value is a fixed value. This value is named as maximum critical value and it is simply $(T_{null}(t), 0, 05)$ (It is shown with dashed red line in Figure 4.)

7. The observed curve is compared with pointwise critical value and maximum critical value at each time point.

8. The average time when $T_{maxobservation}(t)$ is smaller than $T_{null}(t)$ is calculated and this is the related p value . The p_{value} is as follows:

$$p_{value} = mean[T_{maxobservation}(t) < T_{null}(t)] \tag{9}$$

   If p_{value} is smaller than the desired significance level (generally as 0.05), the null hypothesis that is given above is rejected against alternative hypothesis.

9. In order to calculate a related p value at any time, original data are compared with $T_{nullvalues}(t)$ and the mean of $T_{nullvalues}$ that are bigger than the original observations is calculated. In a way, p value for each time point is calculated as,

$$p_t = mean[T_{observation}(t) < T_{nullvalues}(t)] \tag{10}$$

The study of Cox and Lee (2008) and Yaree (2011) are benefited from in forming these steps. The test created this way is also called pointwise t-test.

Statistical package that is formed on the basis of these steps are created by modifying Matlab codes of Ramsay (2013). The program can be found in author's personal web site Keser and Deveci (2013). Data can be smoothed to obtaining observed curve points and pointwise critical value points in order to interpret graphics more easily. Similar to obtaining individual functions, Fourier basis and other basis functions such as wavelets, B-Splines or polynomials may be used for smoothing.

2.2.2 Functional t-tests based on a randomization method of Westfall and Young

Another pointwise procedure used for comparing the mean curves and detect the regions that the differences occur is proposed by Cox and Lee (2008).

Ramsay and Silverman (2005) and Ramsay et al. (2009) didn't take account of the relationship between evaluation points. But with smooth functional data, the events of rejection of $H_0(t_i)$ and $H_0(t_{i+j})$ are highly positively correlated when t_i and t_{i+j} are close. Thus Cox and Lee (2008) suggested an approach that accounts for this correlation. So they use Westfall-Young randomization method, for which Westfall (2005) notes that it can 'account for spatiotemporal correlations as well as non-normal distributional characteristics'. This verified for functional data by Cox and Lee (2008).

In order to avoid the multiplicity problem, the individual p-values are adjusted using Westfall-Young method. The Westfall-Young method is a step-down resampling method. In other words, the testing begins with the first ordered hypothesis (corresponding to the smallest unadjusted p-values) and stop at the first non-rejection. The t-test procedure using the Westfall Young randomization method can be summarized with these basic steps:

1. First, we perform the univariate t-tests ($Eq.2.2$) at every evaluation points (here time points) and obtain the unadjusted p-values from original data. The unadjusted p-values ordered from min to max ($p_{r_1} \leq \ldots \leq p_{r_k}$) as in Holm (1979)'s method.

2. Initialize counting variables $C_i = 0$, $i = 1, 2, \ldots, k$

3. Randomly permute data between the two populations and call the resulting data set as a randomized data set. The p-values computed from a randomized data set are denoted by p^*. Put the p^* values in the same order as the sorted p^*- values for the original data set. Note that the sequence r_j is fixed throughout the simulation. Thus the $p^*_{r_j}$ will not have the same monotonicity as the original p values $p_{(j)}$.

4. Define the successive minima
   $q^*_k = p^*_{r_k}$
   $q^*_{k-1} = min(q^*_k, p^*_{r_{k-1}})$
   $\vdots$
   $q^*_1 = min(q^*_2, p^*_{r_1})$

5. If $q^*_i \leq p_{(i)}$, then $C_i \leftarrow C_{i+1}$

6. Repeat Step (3-5) N times. This means randomly permute observations N times. The adjusted p-value is computed as $\widetilde{p}^N_i = C_i/N$. So the adjusted p-value is the proportion of q^*_i less than or equal to $p_{(i)}$, with an additional constraint given in Step 7.

7. Enforce monotonicity using successive maximization:
   $\widetilde{p}^N_1 = \widetilde{p}^N_1$
   $\widetilde{p}^N_2 \leftarrow max(\widetilde{p}^N_1, \widetilde{p}^N_2)$
   $\vdots$
   $\widetilde{p}^N_k \leftarrow max(\widetilde{p}^N_{k-1}, \widetilde{p}^N_k)$
   Once monotonicity has been enforced, the simulation based estimates p^N_j are reasonably approximations of the actual values $\widetilde{p}_j$, provided N is sufficiently large. $N \geq 10000$ is recommended.

8. The null hypotheses are rejected when the adjusted p-values ($\widetilde{p}^N_i$) are smaller than the desired significance level α (generally as 0.05) $\widetilde{p}^N_i < \alpha$. Also, this method controls the familywise error rate ($FWER$), the probability of committing at least one Type 1 error, to be less than or equal to α.

Westfall and Young (1993), Cox and Lee (2008) and Vsevolozhskaya et al. (2013) are benefited from while forming these steps.

New approaches are developed as alternatives to pointwise t-test. Vsevolozhskaya et al. (2013) propose an alternative procedure for identifying regions of significant

difference in the functional domain. Their procedure is based on a region-wise test and application of a combining function along with the closure multiplicity adjustment principle. Lee (2005), Zhang et al. (2010), Cheng et al. (2010) and Coffey and Hinde (2011) can be read for functional t-test practices and alternative approaches.

3 An Application To Humidity Data:Turkey Case

In this study, humidity differences between coastal areas and hinterland which are resulted from geographical location and landform in Turkey are analyzed. The factors that affect humidity and the effects of geographical location and landform of Turkey on humidity are briefly introduced in this section.

3.1 Preliminary Information

Water that is evaporated from water mass on earth cause dampening in atmosphere. Water vapor in the atmosphere is called air humidity. The amount of water vapor in the atmosphere, which is a significant heat factor, changes according to time and place. The basic factors that affect distribution of humidity in the atmosphere can be summarized as evaporation, heat, altitude and pressure.

Resource of the humidity in the atmosphere is water mass on earth. Evaporation increases in parallel with the increase in heat, lack of humidity, water surface, wind and decrease in pressure. On the other hand, as mentioned before, dehumidification capacity of air is high when the temperature is high, so evaporation increases; and it decreases when temperature is low. Water vapor which is a heavy gas cannot rise very much because of gravity. It rains as a result of condensation and as it becomes colder when it rises, dehumidification of air and accordingly humidity decreases. Finally, descending air movement in high pressure areas prevents evaporation, because increase in the intensity of descending air prevents rising of water vapor. In low pressure area, rising air is less dense; evaporation is easier 9 (2013). So, it is expected that distribution of humidity decreases from seas to land, low area to high areas and from equator to the poles.

In line with this information, when geographic location of Turkey and its landforms are taken into consideration, it is expected that humidity effect in coastal areas and hinterland differs. Turkey is situated between 36°-42° North latitude, 26°-45° East longitude. According to this, Turkey is located in the temperate zone where four seasons clearly occur. Mathematical location is not the only effective factor on climate, temperature, rainfall and humidity. Because of the landforms, heights and seas around the country, climate seriously differs from one region to another. Especially position of regions in terms of seas, location of mountains in terms of seas and their heights have significant effects on humidity. Especially coastal areas in the country have much higher humidity ratio than the inner regions. On the other hand, North Anatolia Mountains and Toros Mountains prevent the effects of seas from entering inner lands. The effect of seas and

Figure 1: Turkey Map

ratio of humidity decrease from coastal areas to hinterland. Similarly, mean elevation of Turkey is high. Elevation increases from west to east and this is why temperature falls in east and humidity decreases in parallel with the increase in height.

3.2 Analysis and Results

In this research, the effect of geographical formations and climate on humidity is explained by comparing coastal area and hinterland humidity mean functions statistically. In our case, it is expected that during most of the year, humidity mean curves moves differently. Functional t-test will be used for this purpose. Year of 2010 will be analyzed in detail and comparisons for 2000-2010 will also be presented visually and summarized. Data used in the study is obtained from Head Office of Meteorology 28 (2013). Before functional t-tests, individual humidity functions and mean functions are analyzed.

In this study, 35 cities which has complete daily mean humidity values for 2000-2010 are chosen. There are many stations and years with missing data. Only stations with complete data are included in this study since some of excluded stations are newly established and some of them have technical data collection problems. Missing data imputation for those stations and years was not a preferable choice because of the high missing value rate.

3.2.1 Obtaining Individual Humidity Functions and Mean Humidity Functions for Coastal Areas and Hinterland

Individual humidity functions of these 35 cities were obtained first with basis function approach by using Fourier series for long term period 2000-2010, because Fourier series are suitable to show the seasonality for long time and each season is observed multiple times. Individual humidity functions of these 35 cities were obtained first with basis function approach by using Fourier series for long term period 2000-2010, because Fourier series are suitable to show the seasonality for long time and each season is observed multiple times.

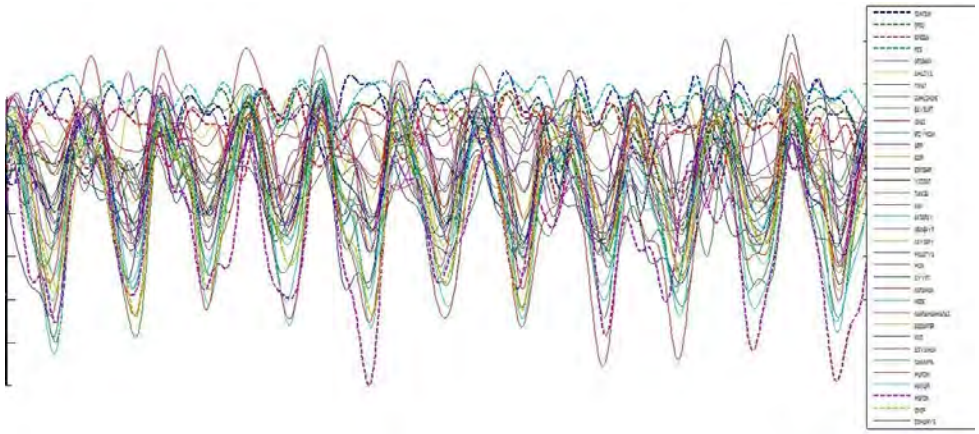

Figure 2: Individual Humidity Functions Fourier basis ($K = 63$)

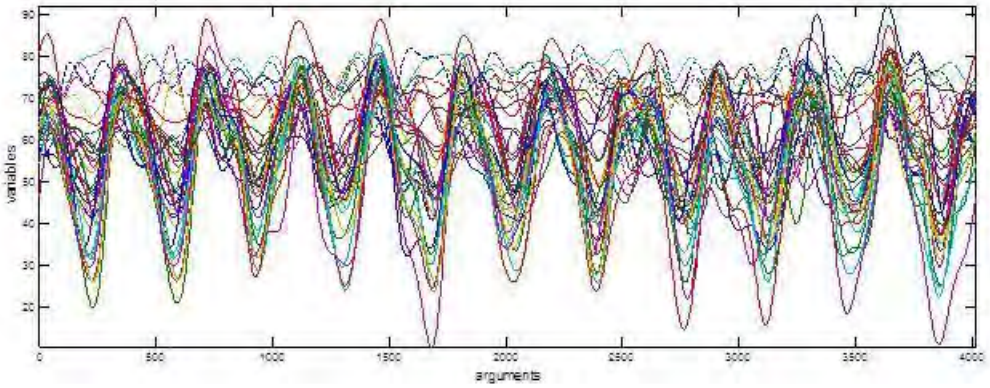

Figure 3: Individual Humidity Functions Fourier basis ($K = 33$)

We first use 63 basis functions, as the square root of evaluation points (4015) is approximately 63. In Figure 2 and 3, cities in coastal areas are presented with dashed lines while hinterland cities are presented with solid lines and they are tagged on the right side of the figures. Or if 33 basis functions (figure 3) are used, 3 per year, we can see the seasonality and the difference between the coastal and hinterlands are like in 63 basis functions. If we want to see the seasonality we suggest using 3 basis functions per year. After a lot of trial and error we have concluded that number of time (evaluation) points and basis functions (K) are closely related. For example, if the aim is to see the seasonal change in daily data for one year, using three basis functions, which includes only the first sine-cosine pair, seems to be appropriate. However, three basis functions are not enough to see the change in daily data for 11 years since data would be over-smoothed. Therefore, the number of time points and the aim of the study should be considered together. Using fewer basis functions may lead to over-smoothing which ends up missing some important changes in data. Since deciding the number of basis functions by trial and error is really cumbersome, using roughness penalty approach after deciding a reasonable number of basis functions (for example three) seems more feasible. If anyone wants to see the daily changes then the number of basis functions should be increased.

Also, from the figures we can see Fourier series are suitable and flexible for long term periodical data. We try long term display also with B-Splines too. If we use a B-Spline to show the seasonality changes between the hinterland and coastal areas, we should use roughness penalty approach to see the changes and smoothing the data. But if we want to see the eleven years cycle explicitly than the smoothing parameter should be increased, because the number of observation points is too high, e.g. 4015. In both basis function approach the eleven years cycle can be seen explicitly.

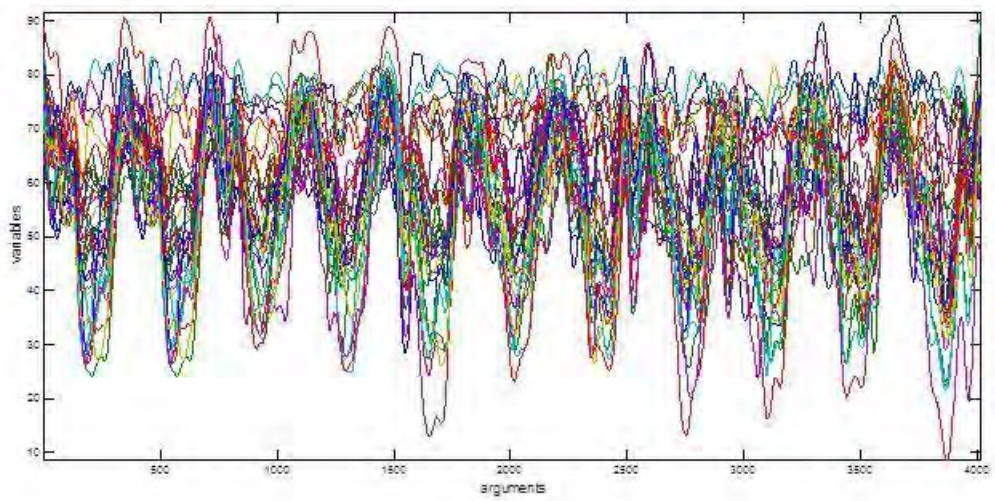

Figure 4: Individual Humidity Functions B-Spline basis ($lambda = 10^2$)

Because daily data are typically noisy and so the smoothing process might lose important information, if we want to see the daily difference between the coastal areas and hinterland or the daily variation, we should analyze data yearly or monthly in both B-Spline and Fourier approaches.

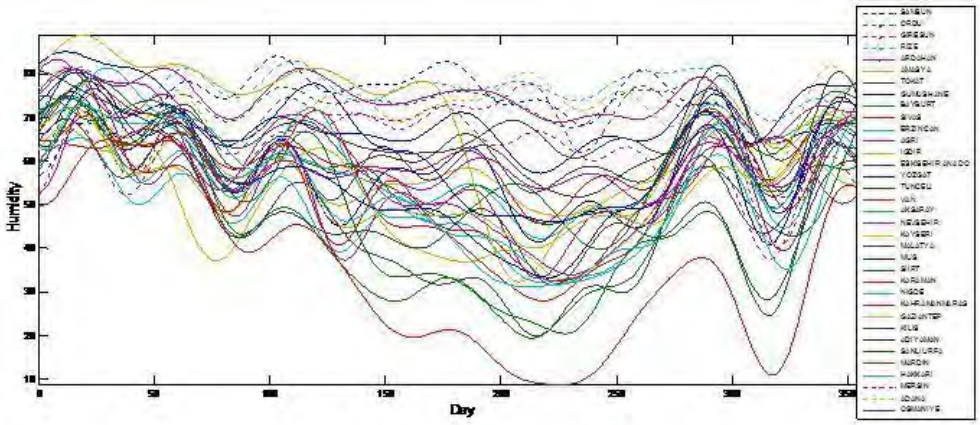

Figure 5: Individual Humidity Functions Fourier basis ($K = 19$)

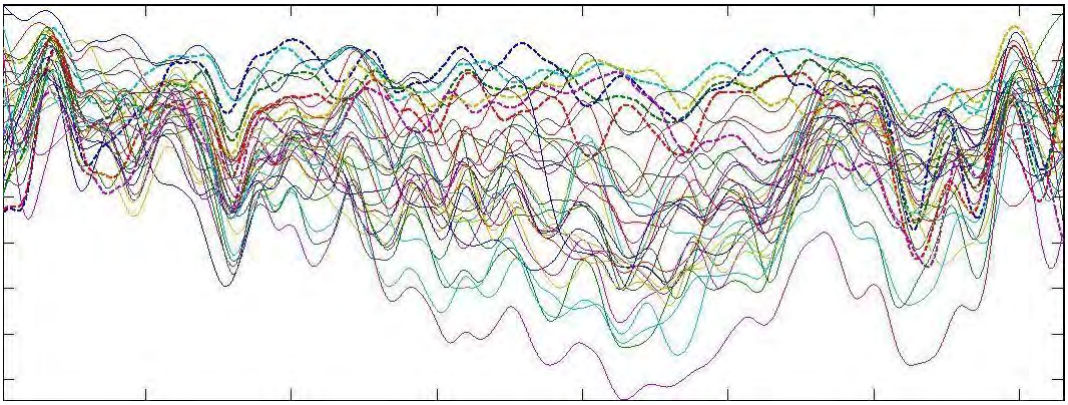

Figure 6: Individual Humidity Functions B-Spline basis ($lambda = 10^2$)

Individual humidity functions of these 35 cities were obtained with basis approach both with Fourier series and B-Splines. We use 19 basis functions per year, as the square root of observation number (365) is approximately 19 in Fourier basis and we use the least squares method to estimate the coefficients. When we use B-Spline approach for smoothing, we must use 367 basis functions and roughness penalty approach to estimate the coefficients. We use the generalized cross validation (GCV) method for smoothing.

In both figures when individual humidity functions of hinterland regions are analyzed, it was seen that when temperature increases, in other words, when summer months start, humidity curves of hinterlands, which are relatively higher and far from coastal effects decrease rapidly as expected and at other times, they follow a little higher course. Here, the city that has the lowest humidity value is the city of Mardin, whose average height is

1082 m and situated in Southeastern Anatolia Region, which is hinterland. On the other hand, when individual curves are analyzed, it is possible to see that differences between coastal areas and hinterlands increase in parallel with the increase in height and being away from coastal areas. All of the coastal areas in our study have mostly high values except winter months during which their values are very close to one another. It can be seen that in hinterlands and high areas, amplitude variation is more.

Both of the approaches show us the differences explicitly. However, if one of them is to be chosen, sum of squares of error may be examined. We suggest to use B-Spline approach especially for one year's data. There are some methods to select a smoothing parameter. The famous is the generalized cross validation. But in applications mostly it is preferred to select the smoothing parameter subjectively and smaller than the GCV value.

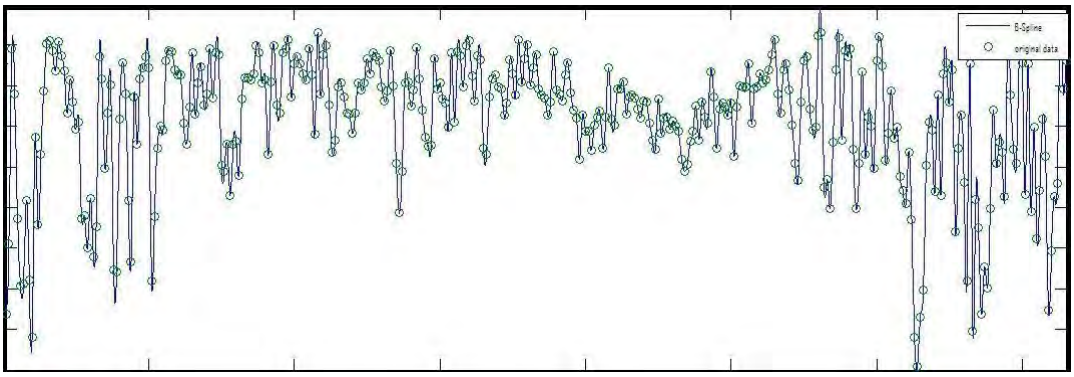

Figure 7: Individual Humidity Functions B-Spline Basis and Original Data

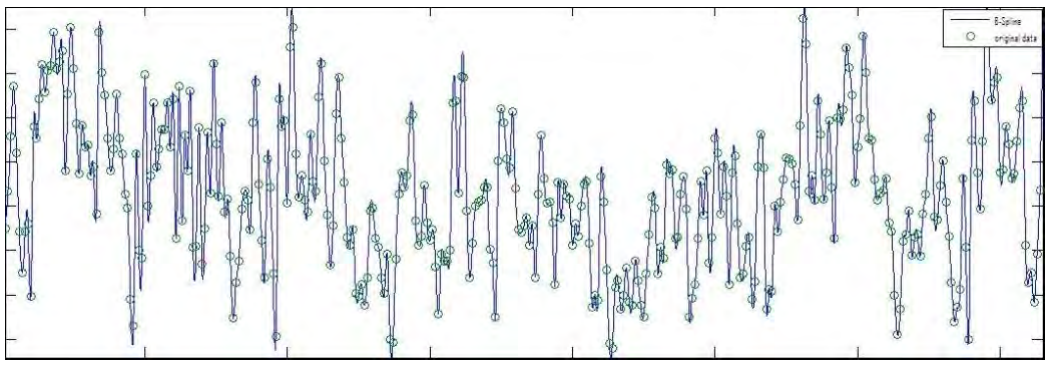

Figure 8: Individual Humidity Functions B-Spline Basis and Original Data

As to show the quality of fit to the original time series data, individual functions for two cities (one from hinterland, and one from coastal area) are plotted in Figure 7

and 8 respectively along with their original data. If the main purpose is increasing the quality of fit, then smoothing should be limited. If the main purpose is interpreting the seasonality or the main variation modes, then smoothing may be used more flexibly.

In order to compare coastal areas and hinterlands, mean curves of both groups are analyzed. In Figure 9, mean humidity functions of coastal areas are presented with a dashed line while hinterlands' mean humidity functions are presented with a solid line. When compared the individual humidity functions, it can be seen well in this figure that, humidity functions of hinterlands are lower than coastal areas most of the year.

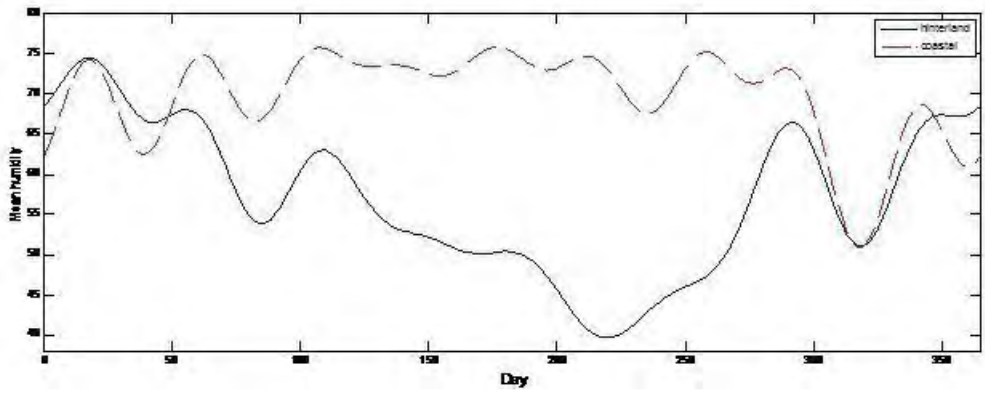

Figure 9: Mean Humidity Functions of Coastal areas and Hinterland

When mean curves are analyzed, it is seen that coastal areas and hinterlands start to differ starting from the 60^{th} day in reverse directions and this difference increased, and especially between the 90^{th} and 270^{th} days, it became more obvious. Towards the end of the year, two regions start to have tendency to move together. During one and a half month in a year and during the last one month, in other words during winter, this difference becomes minimum. It is necessary to test if these findings are statistically verifiable and functional t-tests can be used in order to determine this.

3.2.2 Comparison of Coastal areas and Hinterlands' Mean Functions with Functional t-Test

Functional t-test is used in order to test if the mean of two curve groups are statistically different. We have two curve groups which have different number of curves observed in the same time points. While one group is made of humidity curves of cities in hinterland, the other one stands for humidity curves of coastal cities. As mentioned in Section 2, in order to apply a functional t-test, it is not necessary to have equal number of curves in each group, but each curve should have equal number of observation points. Two methods by Ramsay and Silverman (2005) and Cox and Lee (2008) are applied simultaneously, to improve the validity of results. In this study, our observations in both

coastal areas and hinterlands are analyzed for the year 2010 on the basis of 365 days. We first give the results of Ramsay and Silverman and then the adjusted p-values using the Westfall-Young randomization method by Cox and Lee respectively in Figure 10 and 11.

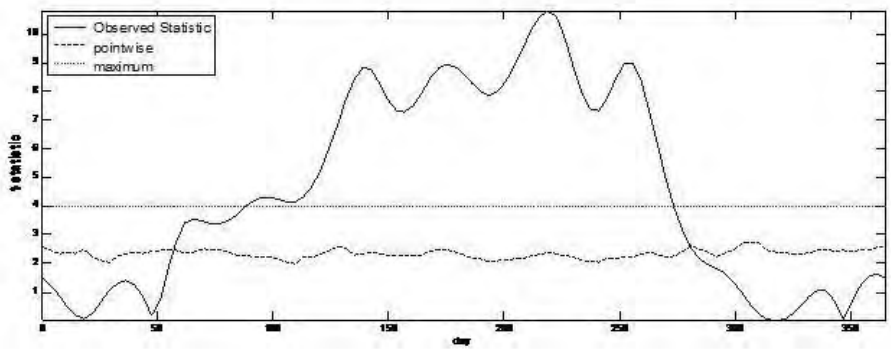

Figure 10: Functional t-test

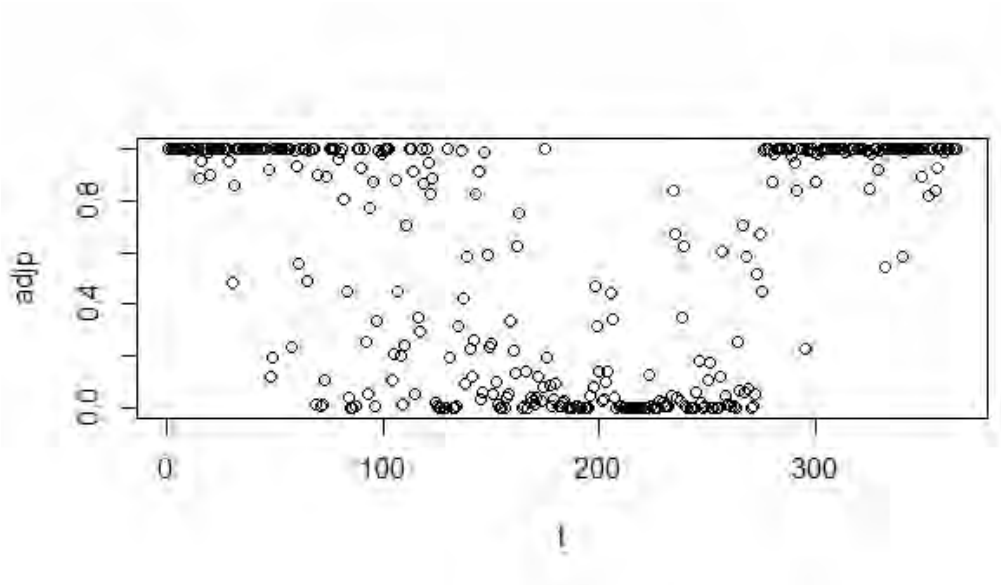

Figure 11: Adjusted p-values

With functional t-test, the time when coastal areas and hinterlands started to differ can be revealed. When pointwise critical value is taken as reference, means of coastal areas and hinterlands tend to differ starting from the 60^{th} day. p value is reported as 0.05 when difference starts. This situation continues until about 280^{th} day. But main strong differences start when it becomes really hot, when conservative reference line (maximum critical value) is taken into consideration, and continues until it becomes to get colder. This situation exists around the period between the 90^{th} and 270^{th} days and this is completely consistent with the expectations of the study. So, it can be concluded that geographical location and structure is effective on humidity.

According to Cox and Lee, the adjusted p-value is found to be less than the critical value of 0.05, meaning that the null hypothesis of "no difference in time points" should be rejected. Hence two approaches validate each other. Obtaining adjusted p-values is easily implemented via the R package multtest Pollard et al. (2011).

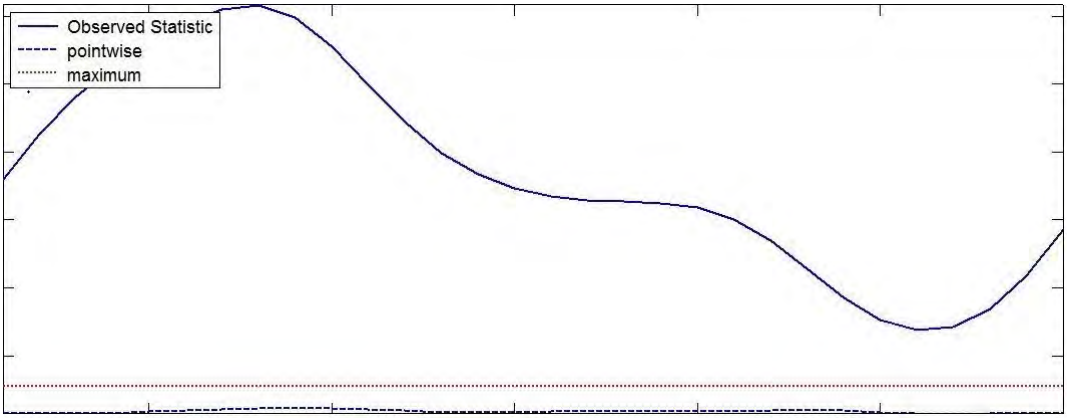

Figure 12: Functional t-test for August 2010

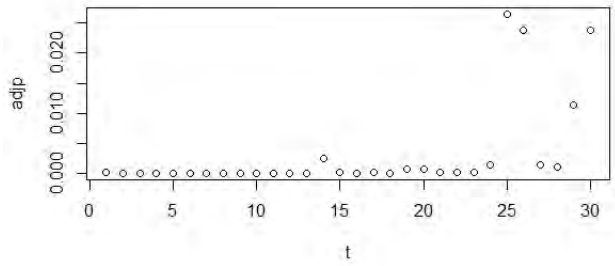

Figure 13: Adjusted p-values for August 2010

When summer months where the differences are likely to occur are examined, observed curves are found to be well above the maximum critical value. In other words in all the time interval there is a strong differences between coastal areas and hinterlands. For all the time interval, the p-values are approximately 0.00. The adjusted p-values confirm this too in Figure 13. The more the observed statistic gets far away from maximum critical value, the more adjusted p-values get closer to 0.00. In other words, the probability of rejecting the null hypothesis will increase.

3.2.3 Analysis of the period between 2000 and 2010

When the data for 2000-2010 years are analyzed together, it is seen that interpretations for 2010 are valid for all of the years. Here, mean functions of 2000-2010 are analyzed through a rainbow plot. Functional data analysis is firstly based on visualization; here, in order to interpret 2000-2010 mean functions on the basis of years, separate rainbow plots for coastal areas and hinterland are used.

In a rainbow plot, colors of curves are put in order according to the colors of a rainbow; the oldest data are presented with red while the newest ones are purple. Rainbow plot is used in this study in order to emphasize the features of data changes in time. Rainbow graphic is especially useful for enlightening the time sequence of data. In a more typical way, it can be used for organizing data according to functional depth or data density. For instance, curves can also be colored according to their distance to median curve or mean curve.

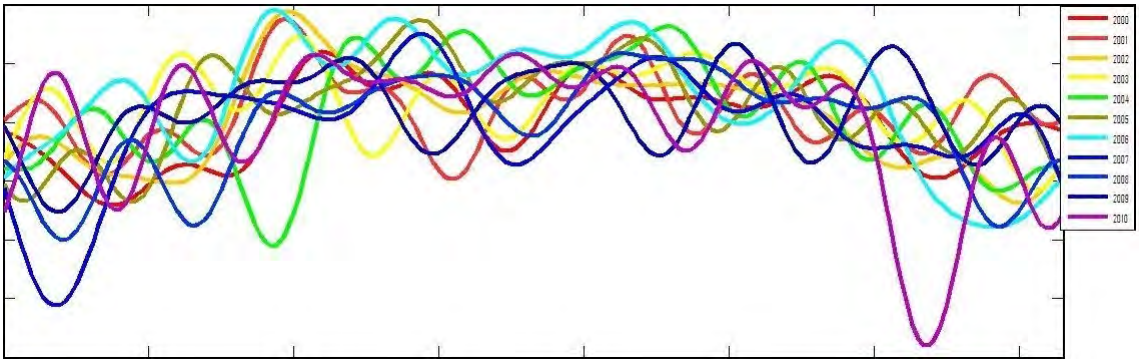

Figure 14: Coastal Areas Rainbow Plots of 2000-2010

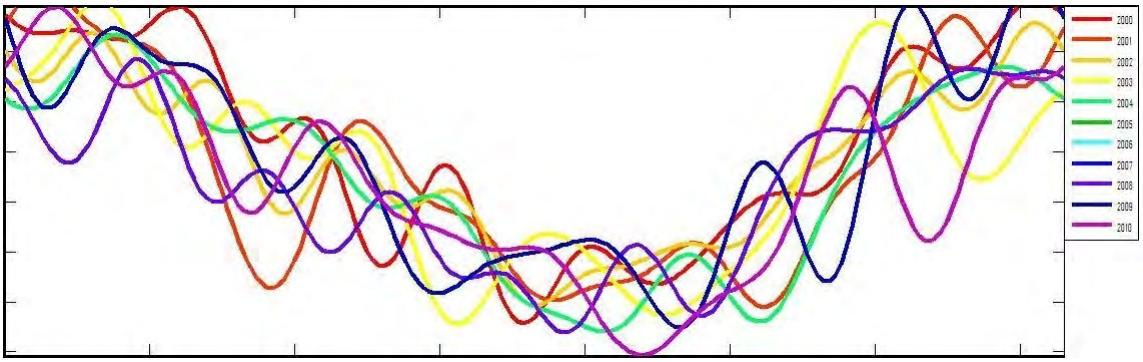

Figure 15: Hinterlands Rainbow Plots of 2000-2010

When Figure 14 and Figure 15 are interpreted together, it is seen that there isn't a regular increase or decrease. In other words, colors don't exist in a rainbow ordering. If there is a regular decrease or increase, the figure will look like a rainbow. Hyndman and Shang (2010) can be examined for details on rainbow plots.

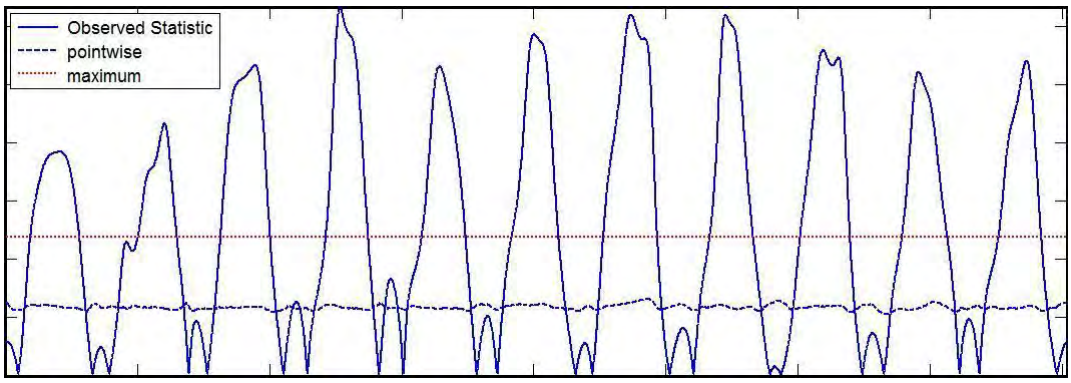

Figure 16: Functional t-test Results of the Years 2000-2010

Before examining all years separately, a functional t test is applied to all years (2000-2010) and overall p-value is found to be less than 0.05. Results of this test can be seen in Figure 16. Results of functional t-test and adjusted p-values that are used for determining mean humidity curves for all of the years for coastal areas and hinterlands are presented graphically in Figure 17-18 respectively.

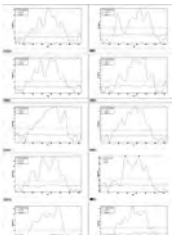

Figure 17: Functional t-test Results of the Years 2000-2009

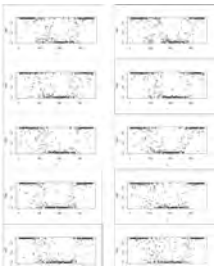

Figure 18: Functional t-test Results of the Years 2000-2009

Observed curves of each year is different from one another. Besides that, when all of the years are analyzed together, it can be seen that although there are one-two days difference in terms of years, coastal areas and hinterlands start to differ from one another starting from the 60^{th} day and this continues until approximately 280^{th} day. p-value is reported to be 0.05 for each when differences start. But the real strong differences are again between 90^{th} and 270^{th} days when conservative reference line, i.e., the maximum critical value is taken into consideration. This can also be seen from the rainbow plot in Figure 19 which all years can be interpreted simultaneously. When all of the years are analyzed together, the same things above can be said for adjusted p-values obtained by Westfall-Young approach too. In the same time interval the p-values are smaller than the significance level α. The both approach by Ramsay and Silverman and Cox and Lee support each other as expected for this application.

Thus when all of the years between 2000 and 2010 are taken into consideration, it can be said that except the coldest winter months, mean humidity functions between coastal areas and hinterlands are statistically different, which is consistent with the expectations of the research. This can be seen in Figure 19 too.

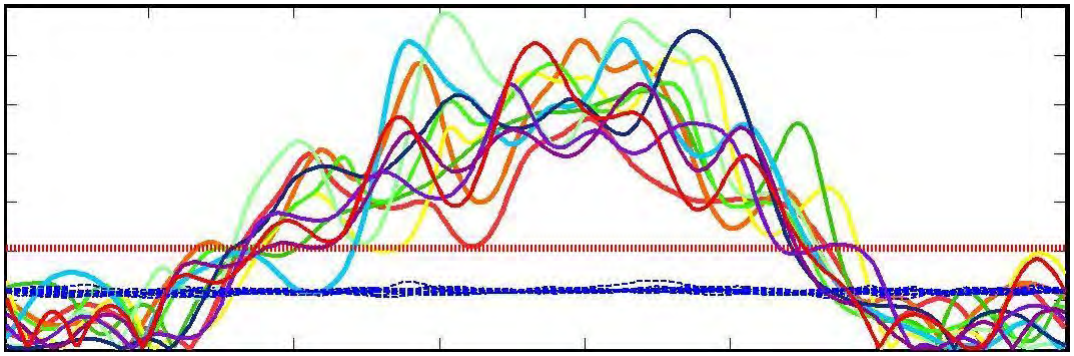

Figure 19: Rainbow plot of the Years 2000-2010

Rainbow plot also shows where the differences start and end for all years as well as if there is a regular increase or decrease. Figure 19 shows no regular increase or decrease.

In this study, it is found out that using either Fourier or B-Spline basis for modeling the differences does not have any considerable effect on p-values. However, smoothing the data makes interpretation easier for both approaches. Therefore, using basis functions for smoothing instead of interpolation can be suggested.

4 Conclusion

In this study, effects of geographical location and landform on humidity curves are investigated while also showing the practicality of functional data analysis. For this purpose,

Turkey, which has a variety of climate and geographical landform, is studied. As Turkey is surrounded by three seas and the elevation increases from west to east, both the geographical situation and the landform affect humidity. According to this, it is expected that humidity curves of coastal areas and hinterlands differ. Although this can be seen when individual humidity curves and mean curves obtained by smoothing data are visually examined, it is also necessary to determine if mean humidity functions of these areas are statistically different. For this aim, humidity curves of 35 cities, whose data for 2000-2010 years are complete, were separated into two groups as coastal areas and hinterlands. In order to see if two curve groups have the same functional curve, in other words, in order to test if humidity mean functions of coastal areas and hinterlands are statistically different, functional t-tests, which were created on the basis of permutation tests and Westfall and Young approach, were used simultaneously to improve the validity of results. As a result of this, although the observed curve that was made of the differences of two curve groups is different in terms of years, it can be easily determined by functional t-test graphics that differences for each year started and ended almost at the same time. The adjusted p-values confirm this too. Except winter months during 2000-2010, mean functions of coastal areas and hinterlands were different from one another at 0.05 significance level. This situation becomes more obvious when temperature increases. One of the significant reasons of this difference is that hinterlands are higher than coastal areas and effects of seas cannot enter the inner lands because of the shape of coasts and mountains. Additionally, rainbow plots are used to identify if there is a regular increase or decrease in the differences or the curves along the years. If a regular movement is detected, the differences may be estimated and/or modeled by some nonparametric approaches.

References

(2013). Meteoroloji Genel Müdürlüğü. URL: http://www.mgm.gov.tr/.

(2013). *Coğrafya Dünyası*. URL: http://www.cografya.gen.tr/.

Ainsworth, L., Routledge, R., and Cao, J. (2011). Functional data analysis in ecosystem research: the decline of oweekeno lake sockeye salmon and wannock river flow. *Journal of Agricultural, Biological, and Environmental Statistics*, 16(2):282–300.

Barra, V. (2004). Analysis of gene expression data using functional principal components. *Computer methods and programs in biomedicine*, 75(1):1–9.

Benko, M. (2004). Functional principal components analysis, implementation and applications. Master's thesis, Humboldt University Center of Applied Statistics and Economics, Berlin.

Besse, P. and Ramsay, J. O. (1986). Principal components analysis of sampled functions. *Psychometrica*, 51(2):285–311.

Cerioli, A., Laurini, F., and Corbellini, A. (2005). Functional cluster analysis of financial time series. *In New Developments in Classification and Data Analysis, eds. Vichi, M., Monari, P., Mignani, S., Montanari, A. Springer-Verlag, Berlin*, pages 333–341.

Cheng, C., Xu, Y., and Gubian, M. (2010). Exploring the mechanism of tonal contraction in taiwan mandarin. *INTERSPEECH 2010*, pages 2010–2013.

Chiou, J. M. and Müller, H. G. (2007). Diagnostics for functional regression via residual processes. *Computational Statistics and Data Analysis*, 51(10):4849–4863.

Coffey, N. and Hinde, J. (2011). Analyzing time-course microarray data using functional data analysis-a review. *Statistical Applications in Genetics and Molecular Biology*, 10(1):1–32.

Cox, D. D. and Lee, J. S. (2008). Pointwise testing with functional data using the westfall-young randomization method. *Biometrika*, 95(3):621–634.

Cuevas, A., Febrero, M., and Fraiman, R. (2004). An anova test for functional data. *Computational Statistics and Data Analysis*, 47:111–122.

Hall, P. and Keilegom, I. V. (2007). Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica*, 17:1511–1531.

Hall, P. and Nasab, H. M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B*, 68(1):109–126.

He, G., Müller, H. G., and Wang, J. L. (2000). Extending correlation and regression from multivariate to functional data. *In Asymptotics in Statistics and Probability, eds. Puri, M.L., VSP, Zeist*, pages 197–210.

He, G., Müller, H. G., and Wang, J. L. (2003). Functional canonical analysis for square integrable stochastic processes. *Journal of Multivariate Analysis*, 85:54–77.

He, G., Müller, H. G., and Wang, J. L. (2004). Methods of canonical analysis of functional data. *Journal of Statistical Planning and Inference*, 122:141–159.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6:65–70.

Hyndman, R. and Shang, H. L. (2010). Rainbow plots, bagplots and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1):29–45.

Ingrassia, S. and Costanzo, G. D. (2005). Functional principal component analysis of financial time series. *In New Developments in Classification and Data Analysis, eds. Vichi, M., Monari, P., Mignani, S., Montanari, A.*, pages 351–358.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society Series B*, 64(3):411–432.

Kaziska, D. M. (2011). Functional analysis of variance, discriminant analysis, and clustering in a manifold of elastic curves. *Communications in Statistics-Theory and Methods*, 40:2487–2499.

Keser, I. K. (2010). Ege bölgesi yağış verilerinin fonksiyonel veri analizi ile incelenmesi. *Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 25(1):40–66.

Keser, I. K. and Deveci, I. K. (2013). FDA Package. URL: http://people.deu.edu.tr/istem.koymen/fda.html.

Kupresanin, A. M. (2008). *Topics In Functional Canonical Correlation And Regression.* PhD thesis, Arizona State University.

Lee, J. S. (2005). *Aspects of Functional Data Inference and Its Applications.* PhD thesis,

Houston, Texas.

Leurgans, S. E., Moyeed, R. A., and Silverman, B. W. (1993). Canonical correlation analysis when data are curves. *Journal of the Royal Statistical Society Series B*, 55(3):725–740.

Lober, E. M. and Villa, C. (2004). Functional principal component analysis of the yield curve.

Pollard, K. S., Gilbert, H. N., Ge, Y., Taylor, S., and Dudoit, S. (2011). *Multtest: resampling-based multiple hypothesis testing. R* package version 2.10.0.

Ramsay, J. O. (2013). FDA Package, URL: http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html.

Ramsay, J. O. and Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B (Methodological)*, 53(3):539–572.

Ramsay, J. O., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer, New-York.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer-Verlag, NewYork, 2nd edition.

Ratcliffe, S. J., Leader, L. R., and Heller, G. Z. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data i: Functional regression. *Statistics In Medicine*, 21(8):1103–1114.

Shen, Q. and Faraway, J. J. (2004). An f test for linear models with functional responses. *Statistica Sinica*, 14:1239–1257.

Ullah, S. and Finch, C. F. (2013). *Applications of functional data analysis: A systematic review*. Medical Research Methodology. URL: http://www.biomedcentral.com/1471-2288/13/43.

Vsevolozhskaya, O. A., Greenwood, M. C., Bellante, G. J., Powell, S. L., Lawrence, R. L., and Repasky, K. S. (2013). Combining functions and the closure principle for performing follow-up tests in functional analysis of variance. *Computational Statistics and Data Analysis*, 67:175–184.

Westfall, P. (2005). Comment on a paper by y. benjamini and d. yekutieli. *J. Am. Statist. Assoc.*, 100:85–89.

Westfall, P. H. and Young, S. S. (1993). *Resampling Based Multiple Sampling: Examples and Methods for p-value Adjustment*. New-York:Wiley.

Yaree, K. (2011). Functional data analysis with application to ms and cervical vertebrae data. Master's thesis, Edmonton, Alberta.

Zhang, C., Peng, H., and Zhang, J. T. (2010). Two samples tests for functional data. *Communications in Statistics - Theory and Methods*, 39(4):559–578.